Small Area Estimation of Health Insurance Coverage in 2007

Mark Bauder and Donald Luery
Small Area Methods Branch
Data Integration Division
U.S. Census Bureau

1 Introduction

The Small Area Health Insurance Estimates program at the U.S. Census Bureau produces estimates of numbers and proportions of those with and without health insurance coverage for demographic groups within states and counties. These demographic groups are defined by age, sex and income, and also for states by race and ethnicity. Income groups are defined in terms of income-to-poverty ratio (IPR), which is the family income divided by the Federal Poverty Level.

For 2007, SAHIE publishes estimates for states of the numbers and proportions insured and uninsured for the following domains.

(1) The full cross classification of

- 4 age categories: 0-64, 18-64, 40-64, 50-64
- 4 race/ethnicity categories: all races, Hispanic, White not Hispanic, Black not Hispanic
- 3 sex categories: all sexes, male, female
- 3 income groups: all income, 0-200% IPR, 0-250% IPR.

(2) Age under 19 in IPR 0-200%.

SAHIE publishes estimates for counties for the following domains.

(1) The full cross classification of

- 3 age categories: 0-64, 18-64, 40-64
- 3 sex categories: all sexes, male, female
- 2 income groups: all income, and either 0-200% IPR or 0-250% IPR, depending on the state

(2) Age under 19 in IPR 0-200%.

The choice of domains is motivated by the needs of one of the SAHIE program's sponsors, the Centers for Disease Control and Prevention (CDC). The CDC has cancer screening programs, for which the eligible population is low-income, uninsured women in specified age groups (SAHIE Team 2008). In addition, the under age 19 low-income category is relevant to the Children's Health Insurance Program (CHIP). Because the SAHIE models produce estimates for disjoint groups covering virtually everyone under age 65, we released estimates for men and children as well as women, and for other aggregates of possible interest.

In the sections to follow, we describe in detail the models used to produce the SAHIE estimates.

2 Overview of SAHIE modeling

In order to obtain estimates for the domains described above, we model disjoint groups at a level fine enough that we can obtain the estimates needed, perhaps by aggregation. We did not think it feasible to estimate domains with only age 18, and for counties we did not think it feasible to estimate domains for only income 200-250% IPR. For these reasons, we estimate the model for two sets of data for states, and three sets of data for counties. For states, we model the full cross-classification of

- race/ethnicity: Hispanic, White not Hispanic, Black not Hispanic, Other not Hispanic,
- sex: male, female,
- IPR: 0-200%, 200-250%, above 250%

with two sets of age categories

(1) ages: 0-17, 18-39, 40-49, 50-64
(2) ages: 0-18, 19-39, 40-49, 50-64 .

For counties, for both sexes by

(1) ages: 0-17, 18-39, 40-64, IPRs: 0-200%, above 200%
(2) ages: 0-18, 19-39, 40-64, IPRs: 0-200%, above 200%
(3) ages: 0-17, 18-39, 40-64, IPRs: 0-250%, above 250%.

We use an area-level model (Rao 2003) in the sense that we model survey estimates at some aggregate level, rather than individual survey responses, and also use other data at aggregate levels. Parts of our model are similar to the well-known Fay-Herriot model (Fay and Herriot 1979). However, our model differs from it in several ways.

(1) We model two quantities instead of one:
   - the number of people in IPR categories, and
   - the proportion of people within the IPR categories with insurance.

(2) We model estimates from Census 2000 Sample Data (hereafter, Census 2000 estimates) and administrative record data as random, rather than using them as fixed predictors in a regression model.

We formulate the model in a fully Bayesian framework. We use a Bayesian model approach because our model is complex enough that estimation in the frequentist framework would be difficult. We model two direct estimates because survey estimates for the numbers in the income groups as well as the numbers with health insurance coverage are not reliable enough at the level we model.

2.1 Modeling auxiliary data

In the standard Fay-Herriot model, there are survey estimates, $\hat{\theta}_i$ of the variable of interest, θ_i and covariates in the form of auxiliary data, $\mathbf{A}_i = (A_{i1}, \ldots, A_{ip})^T$. Then

$$\begin{aligned} \hat{\theta}_i &= \theta_i + e_i \\ \theta_i &= \mathbf{A}_i^T \beta + u_i \end{aligned}$$

where the sampling errors e_i and model errors or random effects u_i are independent, normal, with mean zero. Here, the covariates $\mathbf{A}_i$ are treated as fixed predictors.

Fisher and Gee (Fisher 2003 and Fisher and Gee 2004) proposed an alternative in the context of estimating poverty. In their "errors-in-variables" research for the Small Area Income and Poverty Estimates program, they treated the covariates as measures of the quantity of interest θ (log poverty in their example), that are possibly biased and have random error. In their "error-in-variables" model

$$\begin{aligned} \hat{\theta}_i &= \theta_i + e_i \\ A_{ij} &= b_j + c_j \theta_i + u_{ij} \quad j = 1, \ldots, p \\ \theta_i &\sim \mathcal{N}(\mu, v^{\theta}) \end{aligned}$$

where the b_j and c_j are "bias" parameters to be estimated. A feature of this model is that the influence of $\hat{\theta}_i$ and the A_{ij} on the estimate of θ_i can vary observation to observation, depending on their variances.

This approach was extended to small area estimates of insurance coverage in Fisher, O'Hara, and Riesz (2006). In the SAHIE models, the covariates $\mathbf{z}_i^T = \left(\mathbf{x}_i^T, \mathbf{A}_i^T\right)$ include both fixed predictors $\mathbf{x}_i$ and auxiliary data to be modeled $\mathbf{A}_i = (A_{i1}, \ldots, A_{ip})^T$.

Our approach models both the direct survey estimates and the Census 2000 and administrative record data as possibly nonlinear regressions of the θ_i, and the θ_i are modeled by a generalized linear model. We have

$$
\begin{aligned}
\widehat{\theta}_i &= \theta_i + e_i \\
A_{ij} &= h_j\left(\theta_i\right) + u_{ij} \quad j = 1, \ldots, p \\
g\left(\theta_i\right) &= \mathbf{x}_i^T \gamma + v_i
\end{aligned}
$$

where the e_i are sampling errors, the v_i are independent and identically distributed area-specific random effects, and the u_{ij} are random effects associated with the Census or administrative record's data.

3 The data

We use the following data sources for states and counties.

CPS ASEC direct estimates

We have two sets of direct estimates from the Annual Social and Economic Supplement (ASEC) of the Current Population Survey (CPS):

- estimates of the number in each of the three IPR categories in state by age by race/ethnicity by sex categories, and in counties by age and sex
- estimates of the proportion insured in state by age by race/ethnicity by sex by IPR categories, and in counties by age by sex by IPR categories.

We use a three-year average of the direct survey estimates from the 2007, 2008 and 2009 ASEC surveys. These surveys collect income and health insurance coverage information for the years 2006, 2007, and 2008, so that the average estimates are centered around year 2007.

Census 2000 estimates

We use Census 2000 estimates of the number in IPR categories in state by age by race/ethnicity by sex categories, and in county by age by sex by IPR categories.

Internal Revenue Service exemption data

We use the number of IRS exemptions in age by IPR categories in each state and county. The age categories are 0-17, 18-64, and 65+. We do not have actual ages for the IRS data. We use the number of child exemptions as a proxy for age 0-17 and for 0-18.

Supplementary Nutrition Assistance Program data

For each state and county, we use counts of the number of people participating in the Supplemental Nutrition Assistance Program (SNAP, formerly Food Stamps) from the United States Department of Agriculture.

Medicaid/CHIP participation data

We use Medicaid participation records from the Centers for Medicare and Medicaid Services (CMS). States submit their data to the CMS quarterly. Individuals are in the file if Medicaid covered them for at least one day during the quarter. We have Children's Health Insurance Program (CHIP) participation counts from states and counties gathered from a web page of the Centers for Medicare and Medicaid Services (CMS). We combine the Medicaid and CHIP participation data, and we use the combined data for each state and county in cross-classifications of age by sex.

Demographic population estimates

We use demographic intercensal estimates of the resident population from the U.S. Census Bureau's Population Estimates Program. These estimates are published for the nation, states, and counties by age, sex, race, and Hispanic origin. We adjust the total resident population estimates to create population estimates with a universe similar to the CPS ASEC. The CPS ASEC universe includes the civilian non-institutionalized population of the United States and members of the armed forces in the United States living off post or with their families on post. It excludes all other members of the armed forces and treats college students as residing in their parental homes.

See http://www.census.gov/did/www/sahie/methods/inputs/index.html for more information about these data sources.

4 The state model

We use slightly different models for state and county estimates. In the sections below, we describe in detail the two models.

4.1 Notation

We use the following notation.

ARSH (age, race, sex, Hispanic origin) refers to either an age by race by sex (for states) or age by sex (for counties) category.

a indexes state or county by ARSH category.

i indexes IPR category.

$\boldsymbol{S_j}$ denotes the sample size for the j^{th} category.

$\boldsymbol{POP}$ denotes a demographic population estimate.

$\boldsymbol{N}$ denotes a number of people. $\boldsymbol{N_{ai}^{IPR}}$ denotes the number of people in the a^{th} state or county by ARSH and i^{th} IPR category, and $\boldsymbol{N_{ai}^{IC}}$ denotes the number of people with health insurance coverage in the a^{th} state or county by ARSH and i^{th} IPR category. $\boldsymbol{N_{ai}^{UI}} = N_{ai}^{IPR} - N_{ai}^{IC}$ is the number uninsured.

$\boldsymbol{p_{ai}^{IPR}} = N_{ai}^{IPR}/POP_a$ is the proportion among those in the a^{th} state or county by ARSH group who are in the i^{th} IPR category.

$\boldsymbol{p_{ai}^{IC}} = N_{ai}^{IC}/N_{ai}^{IPR}$ is the proportion among those in the state or county by ARSH by IPR category ai who have health insurance coverage.

$\boldsymbol{\alpha}$ denotes a mean parameter, i.e., a parameter that appears in a model for the mean of the Census 2000 or administrative record data.

$\boldsymbol{\lambda}$ denotes a variance parameter, i.e., a parameter that appears in the model for the sampling variance of the CPS ASEC estimates, or in a model for the variance of the Census 2000 or administrative record data.

Hatted variables such as $\widehat{N}_{ai}^{IPR}$ denote direct survey estimates.

Overlines such as $\overline{CEN}$ denote means.

Note that the parameters α and λ typically depend on one or more of the age, race/ethnicity, sex, or IPR categories. We suppress indices that show these dependencies.

The models used for state and county estimates are similar. In each case, the model has two main parts: an income part and an insurance part. In the income part,we model the CPS ASEC estimates of the number or proportion of people in the IPR categories, as well as: the Census 2000 estimates, the IRS exemption counts, and the SNAP participation counts. In the insurance part, we model the CPS ASEC estimates of the proportion insured and the Medicaid/CHIP participation counts. The income part of the model allows us to estimate N_{ai}^{IPR}, the number of people in IPR category i, within state or county by ARSH category a. The insurance part of the model allows us to estimate p_{ai}^{IC}, the proportion insured within state or county by ARSH by IPR category ai. We combine these to estimate the primary quantities of interest, N_{ai}^{IC} and N_{ai}^{UI}, the number insured and the number uninsured, where

$$N_{ai}^{IC} = p_{ai}^{IC} N_{ai}^{IPR}$$
$$N_{ai}^{UI} = (1 - p_{ai}^{IC}) N_{ai}^{IPR}$$

In the sections that follow, we first describe the state model in detail. Then we describe the differences in the county model. Note that for states, we model three IPR categories, so the indices $i = 1, 2, 3$ denote IPR categories: 0-200%, 200-250%, and above 250%. In practice, we fit the model twice, once where the lowest two age categories are 0-17, 18-39, and once where the lowest two age categories are 0-18, 19-39. The model is the same, only the input data are different. Here, we describe the model assuming 0-17 is the lowest age group.

4.2 The income part of the state model

4.2.1 Modeling the CPS ASEC estimates

In the first part of the income model, we model $\hat{N}_{ai}^{IPR}$, the CPS ASEC estimates of the number in IPR category i within state by ARSH category a. We assume that the CPS ASEC estimates are unbiased, normally distributed, and that, conditional on all N_{ai}^{IPR} and parameters, they are independent of each other and of all the other data in the model. Let $v_{ai}^{S,IPR}$ be the sampling variance of the CPS ASEC estimate. Then the model is

$$\hat{N}_{ai}^{IPR} \sim \mathcal{N}\left(N_{ai}^{IPR}, v_{ai}^{S,IPR}\right)$$
$$v_{ai}^{S,IPR} = POP_s^2 \lambda_0 \frac{q_{ai}(1-q_{ai})}{S_s^{\lambda_1}}$$

where POP_s is the demographic population estimate of the state population, $q_{ai} = N_{ai}^{IPR}/POP_s$ and S_s is the state sample size (the number of households sampled in the state). Here, and below, λ_0 and λ_1 are parameters to be estimated. Here, λ_0 differs by age by IPR categories and λ_1 has one value.

Our motivation for the form of the sampling variance is as follows. The CPS estimates are controlled so that they add to the state population. Thus $\hat{q}_{ai} = \hat{N}_{ai}^{IPR}/POP_s$ so that $\hat{N}_{ai}^{IPR} = POP_s \hat{q}_{ai}$. Thus $\text{var}(\hat{N}_{ai}^{IPR}) = POP_s^2 \,\text{var}(\hat{q}_{ai})$. In the ideal case of simple random sampling with a negligible sampling fraction, we would expect the variance of a direct estimate of the proportion q_{ai} to be $q_{ai}(1-q_{ai})/S_s$. We add parameters to this form to account for the ways in which our situation differs from this ideal. The parameter λ_0 represents the design effect due to the clustering and stratification in the CPS sample design. In addition, research for the Small Area Income and Poverty Estimates (SAIPE) program found that the variance of a CPS ASEC direct estimate decreased as $1/\sqrt{S}$ (Fisher and Asher 2000). We generalize this relationship to an arbitrary power of the sample size, λ_1, and estimate its value.

4.2.2 The regression part of the state income model

We have POP_a, the demographic population estimate from the U.S. Census Bureau's Population Estimates Program, which we consider to be known without error. The numbers in the three IPR categories sum to this population. That is $\sum_{i=1}^{3} N_{ai}^{IPR} = POP_a$. Then p_{ai}^{IPR}, the proportion in IPR category i within state by age by race/ethnicity by sex category a, is $p_{ai}^{IPR} = N_{ai}^{IPR}/POP_a$. We assume that p_{ai}^{IPR} follows a three-category logistic model with both fixed and random effects and normal errors. Conditional on parameters, the errors, ε_{ai} are independent, and the state by IPR random effects, r_{si}^{IPR} are correlated within state, but otherwise independent. We have

$$
\begin{aligned}
p_{ai}^{IPR} &= \frac{\exp(\mu_{ai}^{IPR})}{\sum_{i=1}^{3}\exp(\mu_{ai}^{IPR})} && (1)\\
\mu_{ai}^{IPR} &= x_{ai}'\beta^{IPR} + r_{si}^{IPR} + \varepsilon_{ai}\\
\varepsilon_{ai} &\sim \mathcal{N}\left(0, v^{M,IPR}\right)\\
\left(r_{s1}^{IPR}, r_{s2}^{IPR}\right)' &\sim \mathcal{N}\left((0,0)', \Sigma^{\,r,IPR}\right)\\
\Sigma^{\,r,IPR} &= \left[\begin{array}{ll} v_1^{r,IPR} & \rho^{r,IPR}\sqrt{v_1^{r,IPR}v_2^{r,IPR}}\\ \rho^{r,IPR}\sqrt{v_1^{r,IPR}v_2^{r,IPR}} & v_2^{r,IPR} \end{array}\right]\\
x_{a3} &= (0,\ldots,0)' && (2)\\
r_{s3}^{IPR} &= 0 && (3)
\end{aligned}
$$

where $s = s(a)$ is the state for state/ARSH, a. The ε_{ai} are independent of each other and data, conditional on β^{IPR} and $v^{M,IPR}$. The model variance $v^{M,IPR}$ is the same for all a and i, and the elements of $\Sigma^{\,r,IPR}$ are the same for all states, s. We impose the conditions in (2) and (3) for identifiability of β^{IPR} and the r_{si}^{IPR}.

The predictors in the X matrix are:

- main effects for IPR
- two-way interactions between age and IPR, race/ethnicity and IPR, and between sex and IPR
- three-way interactions among age, race/ethnicity, and IPR
- three-way interactions among age, sex, and IPR
- two-way interactions between IPR and each of the following continuous variables:
  - the log of the state population (from demographic population estimates), standardized[1]
  - the logit of the proportion who are Hispanic in the state (from demographic population estimates), standardized
  - a tax nonfiling rate for the state, standardized.

We standardize the continuous predictors in order to put them on the same scale. This allows for a more meaningful comparison of the regression coefficients.

4.2.3 Modeling state Census 2000 estimates, IRS exemptions, and SNAP counts

We model the means of the Census 2000 estimates, the IRS exemptions, and the SNAP counts as functions of N_{ai}^{IPR}, the number of people in IPR category i, within

[1] We standardize variables by subtracting by their mean and dividing by their standard deviation.

state by ARSH, a.

4.2.4 Modeling Census 2000 estimates for states

We model the Census 2000 estimates, CEN_{ai}^{IPR}, of the number of people in state by ARSH by IPR categories. We assume these estimates have means, $\overline{CEN}_{ai}$, that are linear functions of the N_{ai}^{IPR}, and that conditional on parameters and all N_{ai}^{IPR} they are independent of each other and other data. The model is

$$\begin{aligned} CEN_{ai} &\sim \mathcal{N}(\overline{CEN}_{ai}, v_{ai}) \\ \overline{CEN}_{ai} &= \alpha N_{ai}^{IPR} \\ v_{ai} &= \lambda_0 \overline{CEN}_{ai}^{\lambda_1} \end{aligned}$$

Here and below, the α's are parameters to be estimated. In this case, α has two additive factors: one that differs by age by IPR categories and one that differs by race/ethnicity categories, λ_0 differs by age by race/ethnicity categories, and λ_1 takes on one value.

4.2.5 Modeling IRS exemptions for states

We have the number of IRS exemptions by state by three approximate age categories (0-17, 18-64, and 65+) by IPR categories. The age categories are approximate because the number corresponding to the 0-17 category is actually the number of child exemptions. We assume that the numbers of exemptions are normally distributed with a mean that is a linear function of aggregate N_{ai}^{IPR}'s, and that conditional on parameters and all N_{ai}^{IPR}, they are independent of each other and of all the other data in the model. Let t index state by the three age categories. Then

$$\begin{aligned} TAX_{ti} &\sim \mathcal{N}\left(\overline{TAX}_{ti},\ v_{ti}\right) \\ \overline{TAX}_{ti} &= \alpha N_{ti}^{IPR} \\ v_{ti} &= \lambda_0 \overline{TAX}_{ti}^{\lambda_1} \end{aligned}$$

where N_{ti}^{IPR} is the number of people in state by age by IPR category ti. N_{ti}^{IPR} is obtained by summing N_{ai}^{IPR} over the appropriate age, race/ethnicity, and sex categories. The parameters α and λ_0 differ by age by IPR, and λ_1 has one value.

4.2.6 Modeling SNAP participation for states

SN_s is the number of SNAP participants by state. We model the mean, $\overline{SN}_s$ as a linear function of the number of people in the state in the IPR 0-200% category. We use only the lowest IPR category because of the eligibility requirements for SNAP. We assume that the SN_s are normally distributed, and that conditional on parameters and all N_{ai}^{IPR}, they are independent of each other and of all the other data in the model. Let s index state. Then

$$
\begin{aligned}
SN_s &\sim \mathcal{N}\left(\overline{SN}_s, v_s\right) \\
\overline{SN}_s &= \alpha N_{s1}^{IPR} \\
v_s &= \lambda_0 \overline{SN}_s^{\lambda_1}
\end{aligned}
$$

where N_{s1}^{IPR} is the number of people in the state in the IPR $\leq 200\%$ category. The parameters α, λ_0, and λ_1 each take one value.

4.3 The insurance part of the state model

In the insurance part of the model, we model the CPS ASEC estimates of p_{ai}^{IC}, the proportion insured in the state by ARSH by IPR category, and the combined Medicaid/CHIP data. From this part of the model, we obtain estimates of p_{ai}^{IC}, which enables us to estimate our primary quantities of interest, N_{ai}^{IC} and N_{ai}^{UI}, the number insured and the number uninsured in state by age by race/ethnicity by sex by IPR category ai, by $N_{ai}^{IC} = p_{ai}^{IC} N_{ai}^{IPR}$ and $N_{ai}^{UI} = (1 - p_{ai}^{IC}) N_{ai}^{IPR}$.

4.3.1 Modeling the CPS ASEC estimate of the proportion insured for states

We assume that $\hat{p}_{ai}^{IC}$, the CPS ASEC estimates of the proportion insured, are unbiased, normally distributed, and, conditional on parameters and all p_{ai}^{IC}, independent of each other and of all the other data in the model. We have

$$
\begin{aligned}
\hat{p}_{ai}^{IC} &\sim \mathcal{N}(p_{ai}^{IC}, v_{ai}^{S,IC}) && (4)\\
v_{ai}^{S,IC*} &= \lambda_0 \frac{p_{ai}^{IC}(1 - p_{ai}^{IC})}{S_{ai}^{\lambda_1}} && (5)\\
v_{ai}^{S,IC} &= \min(v_{ai}^{S,IC*}, 0.25) && (6)
\end{aligned}
$$

where S_{ai} is the state by ARSH by IPR sample size (the number of people sampled), λ_0 differs by age by IPR category, and λ_1 differs by IPR category.

4.3.2 The regression part of the insurance model for states

We assume that the logit, μ_{ai}^{IC} of the proportion insured, p_{ai}^{IC} follows a normal linear mixed model. The model errors, given parameters are independent with constant variance. Let s index state. There are state by IPR random effects, r_{si}^{IC}, that are correlated within a state and, conditional on parameters, independent otherwise.

Then

$$
\begin{aligned}
p_{ai}^{IC} &= \text{logit}^{-1}\left(\mu_{ai}^{IC}\right) \\
\mu_{ai}^{IC} &= x_{ai}'\beta^{IC} + r_{si}^{IC} + \varepsilon_{ai} \\
\varepsilon_{ai} &\sim \mathcal{N}(0, v^{M,IC}) \\
\left(r_{s1}^{IC}, r_{s2}^{IC}, r_{s3}^{IC}\right)' &\sim \mathcal{N}\left((0,0,0)', \Sigma^{\,r,IC}\right) \\
\Sigma^{\,r,IC} &= \begin{bmatrix} v_1^{r,IC} & \rho_{12}^{IC}\sqrt{v_1^{r,IC}v_2^{r,IC}} & \rho_{13}^{IC}\sqrt{v_1^{r,IC}v_3^{r,IC}} \\ \rho_{12}^{IC}\sqrt{v_1^{r,IC}v_2^{r,IC}} & v_2^{r,IC} & \rho_{23}^{IC}\sqrt{v_2^{r,IC}v_3^{r,IC}} \\ \rho_{13}^{IC}\sqrt{v_1^{r,IC}v_3^{r,IC}} & \rho_{23}^{IC}\sqrt{v_2^{r,IC}v_3^{r,IC}} & v_3^{r,IC} \end{bmatrix}
\end{aligned}
$$

The parameter $v^{M,IC}$ is the same for all a and i. The parameters $v_1^{r,IC}, v_2^{r,IC}, v_3^{r,IC}$, $\rho_{12}^{r,IC}, \rho_{13}^{r,IC}$, and $\rho_{23}^{r,IC}$ are each the same for all a.

The predictors in the X matrix are

- main effects for age, race/ethnicity, sex and IPR
- all two-way interactions among age, race/ethnicity, sex and IPR.

4.3.3 Modeling Medicaid/CHIP enrollees for states

Let MED_m be the number of people enrolled in Medicaid or CHIP in state by age by sex category, m. We assume that the mean, $\overline{MED}_m$, a function of the number insured in the IPR 0-200% category. We use only the IPR 0-200% category because, due to the eligibility requirements of Medicaid and CHIP, most of the people covered by these programs are in that IPR category. For instance, the income threshold for Medicaid eligibility for working parents was above 200% IPR for only five states in 2008. (See www.statehealthfacts.org/index.jsp for tables showing the income thresholds for Medicaid and CHIP eligibility.) We assume that the Medicaid counts MED_m are independent, conditional on all N_{ai}^{IC} and parameters. We have

$$MED_m \sim \mathcal{N}\left(\overline{MED}_m, v_m\right) \tag{7}$$

$$\overline{MED}_m = \gamma_s \alpha N_{m1}^{IC} \tag{8}$$

$$\gamma_s \sim \text{Gamma}\,(\text{mean} = 1, \text{var} = \delta) \tag{9}$$

$$v_m = \lambda_0 \overline{MED}_m^{\,\lambda_1} \tag{10}$$

where $s = s(m)$ is state. N_{m1}^{IC} is obtained by summing N_{i1}^{IC} over the race/ethnicity categories . The parameter α differs by age by sex (with the exception that it takes just one value for ages 0-17), λ_0 differs by age, and λ_1 takes one value. The γ_s's are state level random effects with variance, δ, that is estimated, and are independent given δ. The γ_s's are multiplicative, rather than additive, effects to ensure that the coefficients of N_{m1}^{IC} are always positive, while still allowing the possibility that the γ_s's reduce the coefficient on N_{m1}^{IC}.

5 The county model

The model for county level estimates is similar to that for state level estimates. The following are the major differences. We describe the details in the following sections.

- For counties, we model two rather than three income categories: either
  - IPR 0-200% and IPR above 200%, or
  - IPR 0-250% and IPR above 250%.
- We do not model by race/ethnicity.
- We model three age categories: 0-17, 18-39, 40-64 (or 0-18, 19-39, 40-64), rather than four age categories.
- We model the CPS ASEC estimate of the *proportion* in an income category rather than the number.
- We use nonlinear functions for the expectations of the Census 2000 estimates and the SNAP participation totals. There is a larger range of populations by county, so small nonlinearities are more likely to make a noticeable difference for counties.
- We assume the IRS exemption totals follow a T distribution, rather than normal.
- For the regression part of the the insurance part of the model, we use a number of continuous county-level predictors.
- There are no state or county random effects in the regressions in either the income part or insurance part of the model.

5.1 The income part of the county model

5.1.1 Modeling the CPS ASEC county estimates of proportions in income categories

For counties, p_{ai}^{IPR} is the proportion of those in county by age by sex group a who are in income category i. For counties, there are two IPR groups, so we only model the proportion in one of them. We assume that the CPS ASEC estimates of the proportion in the lowest IPR category, $\hat{p}_{a1}^{IPR}$, are normally distributed, unbiased, and that conditional on all p_{a1}^{IPR} and parameters, are independent of each other and of all the other data in the model. We model the sampling variance $v_a^{S,IPR}$. We have

$$\begin{aligned}
\hat{p}_{a1}^{IPR} &\sim \mathcal{N}(p_{a1}^{IPR}, v_a^{S,IPR}) \\
v_a^{S,IPR^*} &= \lambda_0 \frac{p_{a1}^{IPR}(1-p_{a1}^{IPR})}{S_a^{\lambda_1}} \\
v_a^{S,IPR} &= \min(v_a^{S,IPR^*}, 0.25)
\end{aligned}$$

where S_a is the number of people in sample in county by ARSH group a. Here, λ_0 and λ_1 differ by age.

5.1.2 The regression part of the income model for counties

For counties, we model the proportion, p_{ai}^{IPR}, in a way similar to states but without random effects. We have

$$\begin{aligned}
p_{a1}^{IPR} &= \text{logit}^{-1}(\mu_a^{IPR}) \\
\mu_a^{IPR} &= x_a'\beta^{IPR} + \varepsilon_a \\
\varepsilon_a &\sim \mathcal{N}\left(0, v^{M,IPR}\right).
\end{aligned}$$

the model variance, $v^{M,IPR}$, is a constant. The ε_a are independent of each other and data, conditional on β^{IPR} and $v^{M,IPR}$. The predictors in the X matrix for the income part of the model for counties are:

- main effects for age and sex
- age by sex interactions
- log county population (standardized)
- logit of proportion Hispanic (standardized)
- main effects for states.

5.1.3 Modeling county Census 2000 estimates, IRS exemptions, and SNAP counts

As with states, we model the means of the Census 2000 estimates, the IRS exemptions, and the SNAP counts as functions of the N_{ai}^{IPR}, summed to the appropriate level.

5.1.4 Modeling the Census 2000 estimates for counties

For counties, we model the Census estimates, CEN_{ai}, of numbers in county by ARSH group a and IPR category i as follows.

$$\begin{aligned}
CEN_{ai} &\sim \mathcal{N}(\overline{CEN}_{ai}, v_{ai}) \\
\overline{CEN}_{ai} &= \alpha_0 \left(N_{ai}^{IPR}\right)^{\alpha_1} \\
v_{ai} &= \lambda_0 \overline{CEN}_{ai}^{\lambda_1}
\end{aligned}$$

where CEN_{ai} is the Census estimate. Here, α_0, α_1 and λ_0 differ by age and IPR, λ_1 has one value.

5.1.5 Modeling IRS exemptions for counties

Let t index county by the three tax age categories. For counties, we have

$$\begin{aligned} TAX_{ti} &\sim T\left(\nu, \text{mean} = \overline{TAX}_{ti}, \text{ var} = v_{ti}\right) \\ \overline{TAX}_{ti} &= \alpha_0 N_{ti}^{IPR} \\ v_{ti} &= \lambda_0 \overline{TAX}_{ti}^{\lambda_1} \end{aligned}$$

where T is the t-distribution, parameterized in terms of the degree of freedom parameter, ν, and the mean and variance. The parameter ν is estimated. N_{ti}^{IPR} is obtained by summing N_{ai}^{IPR} over the appropriate age and sex categories. We use a t-distribution here because when we fit the model assuming normality, some residuals were too large to be consistent with the normality assumption. We did not observe this with states. The parameters α, λ_0 and ν differ by the three age by IPR categories. λ_1 differs by the three tax age categories,

5.1.6 Modeling SNAP participation for counties

Let c index state. For SNAP data, we have

$$\begin{aligned} SN_c &\sim \mathcal{N}\left(\overline{SN}_c, v_c\right) \\ \overline{SN}_c &= \alpha_0 \left(N_{c1}^{IPR}\right)^{\alpha_1} \\ v_c &= \lambda_0 \overline{SN}_c^{\lambda_1}. \end{aligned}$$

Note that as with states, we predict SNAP participation from only the lowest IPR category. The parameters $\alpha_0, \alpha_1, \lambda_0$, and λ_1 each take one value.

5.2 The insurance part of the county model

As with states, we obtain estimates of p_{ai}^{IC}, which enables us to estimate our primary quantities of interest, N_{ai}^{IC} and N_{ai}^{UI}, the numbers insured and uninsured in county by age by sex by IPR category ai, by

$$N_{ai}^{IC} = p_{ai}^{IC} N_{ai}^{IPR} \text{ and } N_{ai}^{UI} = (1 - p_{ai}^{IC}) N_{ai}^{IPR}.$$

5.2.1 Modeling the CPS ASEC estimate of the proportion insured for counties

For counties, the model of the CPS ASEC direct estimate, $\hat{p}_{ai}^{IC}$ of the proportion insured, p_{ai}^{IC} is the same as that given above in (4) - (6). For counties, λ_0 differs by age and IPR, and λ_1 differs by IPR.

5.2.2 The regression part of the insurance model for counties

As with states, we assume that the logit of the proportion insured follows a normal linear model. We have

$$
\begin{aligned}
p_{ai}^{IC} &= \text{logit}^{-1}(\mu_{ai}^{IC}) \\
\mu_{ai}^{IC} &= x'_{ai}\beta^{IC} + \varepsilon_{ai} \\
\varepsilon_{ai} &\sim \mathcal{N}(0, v^{M,IC})
\end{aligned}
$$

Here $v^{M,IC}$ is the same for all a, i.

The predictors in the X matrix are:

- age, sex, and IPR main effects and all their two-way interactions
- age by sex by IPR interactions
- the following continuous variables, all standardized as described in footnote 1, and interacted with age, sex, and IPR:
  - log of the population (from demographic population estimates)
  - variance of the log ratios of income to the Federal Poverty Threshold, where income is based on tax records (see Fisher and Turner 2003)
  - logit of the proportion who are Hispanic (from demographic population estimates)
  - logit of the proportion who are non-citizens (from Census 2000)
  - logit of the proportion who are American Indian or Alaskan Native, standardized (from demographic population estimates)
  - logit of the proportion of owner-occupied housing units (from Census 2000)
  - logit of the proportion of households in rural areas (from Census 2000)
  - logit of the ratio of number of employees in retail firms (from County Business Patterns) to the number of people aged 18-64 (from demographic population estimates)
  - logit of the ratio of number of employees in non-retail firms with less than 20 employees (from County Business Patterns) to the number of people aged 18-64 (from demographic population estimates)
  - logit of the ratio of number of employees in non-retail firms with 100 or more employees (from County Business Patterns) to the number of people aged 18-64 (from demographic population estimates).

5.2.3 Modeling Medicaid/CHIP enrollees for counties

The model for Medicaid and CHIP enrollees for counties is the same as that for states, given in (7) - (10).

5.3 Prior distributions

For the Bayesian modeling, we generally use vague priors for the high level parameters. Typically, we choose a mean that is similar to what we expect the posterior mean to be, and a variance large enough so that we do not expect our estimates to be sensitive to this choice. For the regression coefficients β^{IPR} and β^{IC}, we use the (improper) uniform prior over the real numbers of appropriate dimension. For other parameters with support on all of $\mathbb{R}$, we use normal distributions with large variances. We generally use normal distributions with large variances truncated at zero for parameters that must be positive. For the correlation parameters in the random effects in the regressions for the state model, we use uniform priors over the support of the parameters. That support is determined by the requirement that the covariance matrix be positive definite.

6 Model selection

We made many modeling decisions to arrive at the current SAHIE models. In addition to the overall form of the model, these decisions include choices of predictors, and mean and variance functions, and distributions. We describe some of the criteria we used in the next sections.

6.1 Model diagnostics

6.1.1 Standardized residuals

Some choices of mean, variance, and density functions resulted from perceived lack of fit based on diagnostics we use. Our primary model diagnostic is a certain type of standardized residual. For the survey estimates and Census and administrative data that we model, we predict means and variances so that for any data, y, that we model, we can calculate the standardized residual [2]

$$E_{\theta|data}\left[\frac{y - E(y|\,\theta)}{\sqrt{\mathrm{var}(y|\,\theta)}}\right]$$

from the Markov chain Monte Carlo (MCMC) output used to fit the model. See Chib and Greenberg (1995) for an explanation of MCMC. The outer expectation is from the posterior distribution of the parameter θ given the data. For example, for $\hat{N}_{ik}^{IPR}$, the CPS ASEC estimate of the number in an IPR category, the standardized

[2] A common standardized residual is gotten by estimating $E_{\theta|data}(y)$ and $\mathrm{var}_{\theta|y}(y)$ by plugging in point estimates of the parameters involved, and taking $(y - \widehat{E_{\theta|data}}(y))/\sqrt{\widehat{\mathrm{var}_{\theta|data}}(y)}$. We prefer the former type because the variance usually depends on the mean. However, in practice, we see very little difference.

residual is

$$E_{\theta|data}\left[\frac{\hat{N}_{ik}^{IPR}-N_{ik}^{IPR}}{\sqrt{v_{ik}^{S,IPR}}}\right]$$

If the model is correct and y is normally distributed, this standardized residual is distributed as approximately normal(0,1). We look at plots of standardized residuals against various quantities such as the predicted mean, population, predicted variance, and where appropriate, sample size. We also look at boxplots of standardized residuals for different values of categorical variables such as age and IPR, and against quantiles of population, proportions in IPR 0-200%, or proportions insured.

6.1.2 Posterior predictive p-values

Another model diagnostic that we use is the posterior predictive p-value (PPP-value) (Gelman, Meng, and Stern (1996)). A posterior predictive p-value is a measure of how surprising or improbable some function of the data (and possibly parameters) is, under the posterior predictive distribution of that data. Let y represent all of the data and θ represent all of the parameters. A PPP-value is defined as $P_{y^{rep},\theta|y}(T(y^{rep},\theta) \geq T(y,\theta))$ for some function T where the probability is with respect to $p(y^{rep}|\theta)p(\theta|y)$, the joint distribution of a replication of the data, y^{rep}, and θ, conditional on y. Let y_i represent a single data point. We use the functions $T_1(y,\theta) = y_i$ and $T_2(y) = (y_i - E(y_i|\theta))^2$. Thus, the PPP-value corresponding to T_1 is $P_{y^{rep},\theta|y}(y_i^{rep} \geq y_i)$. We refer to this PPP-value as the PPP-value for the mean because many values near 0 or near 1 suggest that means given by the model are generally too low, or too high, respectively. We refer to the PPP-value corresponding to T_2 as the PPP-value for the variance since it measures the surprise in the squared distance between the data and its mean. We compute PPP-values for each of the data sources in the model. We look at plots of PPP-values against various quantities, such as population, posterior means, posterior variances, and sample sizes. We also look at boxplots of PPP-values for different values of categorical variables. Our approach is to use the PPP-values informally. We look for plots that have regions in which many of the PPP-values are near 0 or near 1.

6.2 Selecting predictors for the regression parts of the income and insurance models

In order to select predictors for the income and insurance parts of the model, we look at the posterior means and variances of the regression coefficients. We form an approximate 95% credible interval for the regression coefficient by taking its posterior mean plus or minus two times its posterior standard deviation. Generally speaking, we include a predictor in the model if the approximate 95% credible interval does not include zero.

7 Benchmarking

We benchmark SAHIE estimates of the numbers insured and uninsured in order to make them consistent with a set of national CPS ASEC estimates, and to make county estimates consistent with state estimates. We benchmark state estimates to a relatively small set of national direct estimates of numbers insured and uninsured. We benchmark all possible county estimates to the corresponding state estimates. The benchmarking procedure for counties is a simple proportional adjustment. The procedure for states is more complex.

7.1 State to national benchmarking.

We benchmark the state estimates to CPS ASEC national estimates of insured and uninsured for the following categories:

- IPR 0-250%
- age 0-17, IPR 0-250% (or age 0-18, IPR 0-200%
- age 18-64 (or age 19-64)
- Hispanic
- not Hispanic
- White not Hispanic
- Black not Hispanic.

7.1.1 Methodology for state to national benchmarking

The benchmarking procedure that we use was developed by Luery (1986) in the context of controlling survey weights to control totals. The procedure is as follows. Let B be the number of benchmarks (here, 14), and let $\hat{\mathbf{N}} = (\hat{N}_1, \hat{N}_2, \ldots, \hat{N}_B)'$, be the benchmarks. Let S be the number of small area, or model, estimates, and let $\hat{\mathbf{Y}} = (\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_S)'$. be those estimates. We want to adjust the model estimates so that their sums over states equal the benchmarks. Let b index the benchmarks, let i index the area (here, state by ARSH by IPR by insured/uninsured). Let $\mathbf{X} = (x_{ib})$ be the S x B matrix such that $x_{ib} = 1$ when area i contributes to benchmark b, and 0 otherwise. Then the adjusted estimates $\hat{Y}_i^*$ meet the constraints when $\sum_{i=1}^{S} x_{ib}\hat{Y}_i^* = \hat{N}_b$ for all b.

We want a set of benchmarked estimates that are, in some sense, optimal. Generally, benchmarked estimates are preferable when they are close to the original estimates. We choose to minimize the relative quadratic loss function

$$\sum_{i=1}^{S} (\hat{Y}_i^* - \hat{Y}_i)^2 / \hat{Y}_i \ . \tag{11}$$

That is, we minimize the squared change from the original to the benchmarked estimate, relative to the size of the original estimate. It can be shown that there exists a unique set of $\hat{Y}_i^*$ that sum to the benchmarks and minimize (11). This optimal set of benchmarked estimates, $\hat{\mathbf{Y}}^* = (\hat{Y}_1^*, \hat{Y}_2^*, \ldots, \hat{Y}^*)'$ is given by

$$\hat{\mathbf{Y}}^* = \hat{\mathbf{Y}} + \mathbf{D}(\hat{\mathbf{Y}})\mathbf{X}\mathbf{P}(\hat{\mathbf{N}} - \mathbf{X}^{\mathbf{T}}\hat{\mathbf{Y}}) \tag{12}$$

where $\mathbf{D}(\hat{\mathbf{Y}})$ is a diagonal matrix with the entries of $\hat{\mathbf{Y}}$ along the diagonal and $\mathbf{P} = [\mathbf{X}^{\mathbf{T}}\mathbf{D}(\hat{\mathbf{Y}})\mathbf{X}]^{-\mathbf{1}}$.

For the i^{th} area, this can be written as

$$\hat{Y}_i^* = \hat{Y}_i(1 + \sum_{b=1}^{B} f_b x_{ib}) \tag{13}$$

where the f_b are the B factors given by $\mathbf{F} = (f_b) = \mathbf{P}(\hat{\mathbf{N}} - \mathbf{X}^{\mathbf{T}}\hat{\mathbf{Y}})$. Thus, the choice of the relative quadratic loss function ensures that if two areas i and i' have the same indicators, that is, if $x_{ib} = x_{i'b}$ for all b, then they receive the same proportional change to their estimates, as given in (13).

7.1.2 Variance of state benchmarked estimates

We estimate the models using MCMC methods in which a procedure for generating values from the posterior distribution of all unknown variables is repeated for many iterations. We can obtain an estimate of the variance of the benchmarked estimates by repeating the benchmarking procedure at each iteration of the MCMC, using each time a newly generated set of unbenchmarked estimates. However, the benchmarking totals themselves are estimates, and have some uncertainty. If we treat them as fixed in the benchmarking procedure, we will likely underestimate the uncertainty in the benchmarked estimates.

We address this issue as follows. We have variance estimates for the CPS national estimates we benchmark to. These estimates are large, so they should be close to normally distributed. We approximate the distribution of the benchmark by assuming that it has a posterior distribution that is normal, with mean at the estimate, with the estimated variance. Then, in each iteration of the MCMC, we draw a value for each benchmark total from this approximate distribution. We then do the benchmarking procedure, controlling to these generated totals. In this way, the variability in the benchmarked estimates will come from both the variability of the unbenchmarked estimates and the variability of the benchmark totals, as it should.

7.2 Methodology for county to state benchmarking

We benchmark county estimates so that in each state, the county estimates for insured and uninsured in each age by sex by IPR group sum to the benchmarked state estimates. For each cross-classification of age, sex, and income, we apply an adjustment factor to the county estimates of the number insured and the number uninsured so that the sum of the county estimates equals the state estimate. Let c index counties, j index age by sex categories, i index income categories, and s index states. The adjusted estimate of the numbers insured and uninsured are given by

$$\hat{N}_{cji}^{IC,adjusted} = \frac{\hat{N}_{sji}^{IC}}{\sum_c \hat{N}_{cji}^{IC}} \hat{N}_{cji}^{IC} \qquad \hat{N}_{cji}^{UI,adjusted} = \frac{\hat{N}_{sji}^{UI}}{\sum_c \hat{N}_{cji}^{UI}} \hat{N}_{cji}^{UI}$$

where $\hat{N}_{sji}^{IC}$ and $\hat{N}_{sji}^{UI}$ are the state estimates of the insured and uninsured for age by sex by income categories, and the sums are over the counties, c, in state s.

For variance estimation, in order to take into account the fact that the state estimates have error, we perform the adjustment procedure in each iteration of the MCMC similar to that for state to national benchmarking. In each iteration of the MCMC, we simulate the state control from a normal distribution whose mean is the state estimate and whose variance is the variance of the state estimate.

References

Chib, S. and Greenberg, E. (1995), "Understanding the Metropolis-Hastings Algorithm", *The American Statistician*, 49, 327-335.

Fay, R.E., and Herriot, R.A. (1979), "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data", *Journal of the American Statistical Association*, 74, 269-277.

Fisher, R. (2003), "Errors-In-Variables Model for County Level Poverty Estimation", SAIPE Working Paper, Washington, DC, U.S. Census Bureau.
http://www.census.gov/did/www/saipe/publications/files/tech.report.5.pdf

Fisher, R. and Asher, J. (2000), "Alternate CPS Sampling Variance Structures for Constrained and Unconstrained County Models", SAIPE Technical Report #1, Washington, DC, U.S. Census Bureau.
http://www.census.gov/did/www/saipe/publications/files/tech.report.1.revised.pdf

Fisher, R. and Gee, G. (2004), "Errors-In-Variables County Poverty and Income Models", *2004 American Statistical Association Proceedings of the Section on Government and Social Statistics*.
http://www.census.gov/did/www/saipe/publications/files/
FisherGee2004asa.pdf

Fisher, R., O'Hara, B. and Riesz, S. (2006), "Small Area Estimation of Health Insurance Coverage: State-Level Estimates for Demographic Groups", *2006 American Statistical Association Proceedings of the Section on Government and Social Statistics*.

Fisher, R. and Turner, J. (2003), "Health Insurance Estimates for Counties", *2003 American Statistical Association Proceedings of the Section on Survey Research Methods*.

Gelman, A., Meng, X.-L., and Stern, H. (1996), "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies" (with discussion), *Statistica Sinica*, 6, 733-807.

Luery, D. (1986), "Weighting Survey Data Under Linear Constraints on the Weights", *1986 American Statistical Association Proceedings of the Section on Survey Research Methods*, 325-330.

Rao, J.N.K. (2003), *Small Area Estimation*, New York: Wiley.

Small Area Health Insurance Estimates Team, U.S. Census Bureau (2008), "The

Feasibility of Publishing County-level Estimates of the Number of Women Eligible for the CDCs NBCCEDP",
http://www.census.gov/did/www/sahie/publications/files/cdc_feasibility_report_oct2008.pdf.